# Optimized error analysis for expensive data

Ulli Wolff

27. November 2006

ALPHA Collaboration

$\Rightarrow$ Definition of the problem, basic formulas

$\Rightarrow$ Binning strategy, $\Gamma$-method

$\Rightarrow$ Generalization: many observables

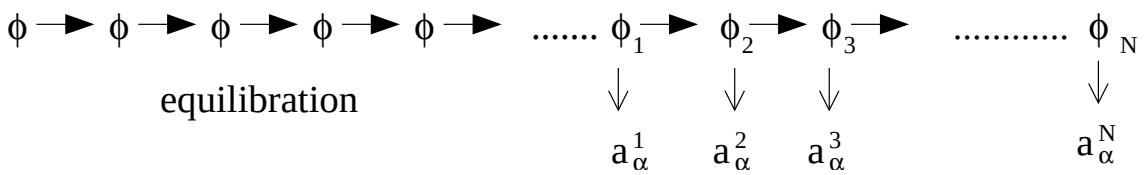
$\Rightarrow$ Generalization: replica simulations, $Q$-value

$\Rightarrow$ Description of software for download

$\Rightarrow$ Sample applications

- all details found in paper hep-lat/0306017 (v4 soon, minor updates)
- these transparencies: `www-com.physik.hu-berlin.de`
- software download: `www-com.physik.hu-berlin.de/ALPHAsoft`

# The problem




- $a^i_\alpha, i = 1, \ldots, N$ successive MC estimates of observables with exact means $A_\alpha$
- recording started $(i = 1)$ after equilibration
- we want to estimate $A_\alpha$ and functions $F = f(A_1, A_2, \ldots)$

$$\text{obvious estimators: } \bar{a}_\alpha = \frac{1}{N}\sum_{i=1}^{N} a^i_\alpha, \quad \bar{F} = f(\bar{a}_1, \bar{a}_2, \ldots)$$

$$\text{covariance matrix: } \sigma^2_{\alpha\beta} = \left\langle (\bar{a}_\alpha - A_\alpha)(\bar{a}_\beta - A_\beta) \right\rangle$$

statistical errors:

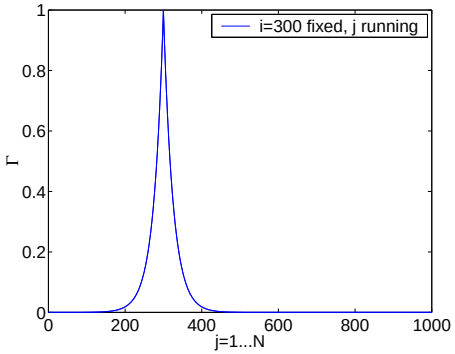
$$\sigma^2_{\alpha\alpha} = \left\langle (\bar{a}_\alpha - A_\alpha)^2 \right\rangle$$

$$\sigma^2_F = \left\langle (\bar{F} - F)^2 \right\rangle \approx \sum_{\alpha\beta} \frac{\partial f}{\partial A_\alpha}\frac{\partial f}{\partial A_\beta}\,\sigma^2_{\alpha\beta}$$

- $\langle \cdots \rangle \leftrightarrow$ infinite ensemble of Markov chains (same algorithm, independent random #)
- simplification for a while: only one $\alpha \longrightarrow$ index dropped (restored later)

$$\sigma^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} \Big\langle (a^i - A)(a^j - A) \Big\rangle = \frac{1}{N^2} \sum_{i,j=1}^{N} \Gamma(j-i) \approx \frac{1}{N} C$$

with $C = \sum_{t=-\infty}^{\infty} \Gamma(t), \quad \Gamma(t) \sim \exp(-|t|/\tau), \quad N \gg \tau$




- $\Big\langle (a^i - A)(a^{i+t} - A) \Big\rangle = \Gamma(t) = \Gamma(-t) =$ autocorrelation function at equilibrium
- $\Gamma(t \neq 0)$ depends on the algorithm, $\Gamma(0)$ is the (static) variance
- $C$ is like a susceptibility, define $2\tau_{\mathrm{int},A} = C/\Gamma(0)$
- only if the whole update obeys detailed balance (and not just balance or stability) one can show that $\Gamma$ is a sum of decaying exponentials with positive coefficients

Our (any) estimator for $F$ is biased unless $F$ is linear:

$$\left\langle \bar{F} - F \right\rangle = \langle f(\bar{a}) - f(A) \rangle \approx \frac{f''(A)}{2} \left\langle (\bar{a} - A)^2 \right\rangle \approx \frac{1}{2N} f''(A) C$$

statistical error:

$$\sigma_F^2 = \left\langle (\bar{F} - F)^2 \right\rangle = \left\langle (f(\bar{a}) - f(A))^2 \right\rangle \approx (f'(A))^2 \left\langle (\bar{a} - A)^2 \right\rangle \approx \frac{1}{N} (f')^2 C$$

systematic bias $\propto C/N$ usually negligible compared to statistical error $\sigma_F \propto \sqrt{C/N}$

$\Rightarrow$ we need to estimate $C$ which will then have itself a statistical error

# Binning strategy

- divide $N = N_B \times B$ measurements into bins of $B$ consecutive measurements each
- form bin-averages $b^k = \frac{1}{B} \sum_{i \in \mathrm{bin}\#k} a^i$
- mean $\bar{b} = \frac{1}{N_B} \sum_{k=1}^{N_B} b^k \equiv \bar{a}$ identical
- one negelects autocorrelations between $b^k$ and uses the standard estimator for the error:

$$\bar{\sigma}^2_{\mathrm{bin}} = \frac{1}{N_B(N_B-1)} \sum_{k=1}^{N_B} (b^k - \bar{b})^2 := \frac{\bar{C}_{\mathrm{bin}}}{N}$$

this implies an estimate $\bar{C}_{\mathrm{bin}}$ for $C$ with bias and statistical error

$$\left\langle \bar{C}_{\mathrm{bin}} - C \right\rangle \sim -\frac{\tau}{B} C \qquad \left\langle (\bar{C}_{\mathrm{bin}} - C)^2 \right\rangle \approx \frac{2}{N_B} C^2 = \frac{2B}{N} C^2$$

- $B$ has to be chosen to compromise between these
- binning $\longrightarrow$ jackknife binning better for estimating $F$, bias and error of $\bar{C}_{\mathrm{jack}}$ the same

# $\Gamma$-strategy

estimate autocorrelation function $\Gamma(t)$ and $C$ explicitly [Sokal], estimator:

$$\bar{\Gamma}(t) = \frac{1}{N-t}\sum_{i=1}^{N-t}(a^i - \bar{a})(a^{i+t} - \bar{a}), \qquad \langle\bar{\Gamma}(t) - \Gamma(t)\rangle \simeq -\frac{C}{N}$$

giving

$$\bar{C}(W) = \bar{\Gamma}(0) + 2\sum_{t=1}^{W}\bar{\Gamma}(t)$$

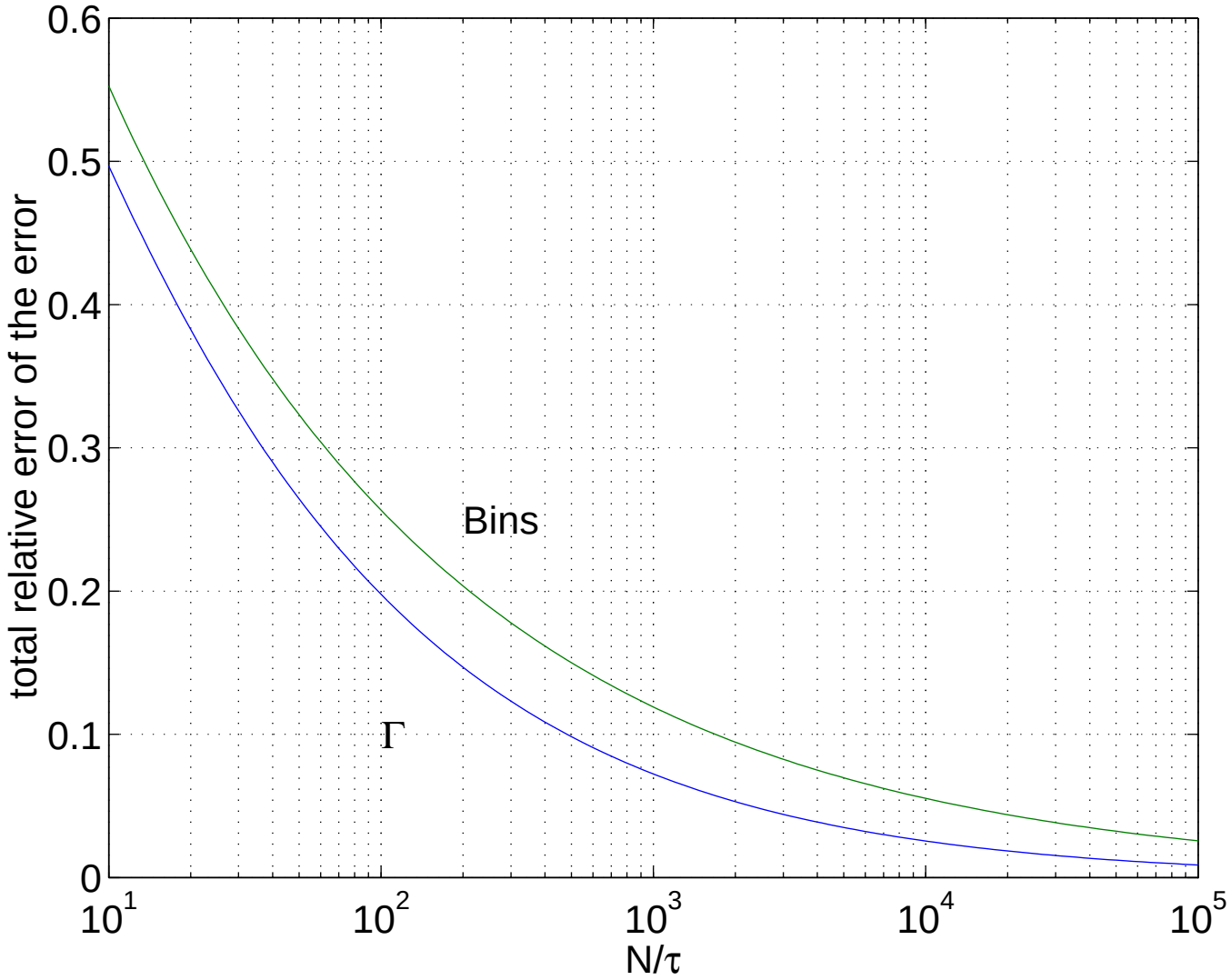
with bias and statistical error

$$\left\langle \bar{C} - C \right\rangle \sim -\exp(-W/\tau)\, C \qquad \left\langle (\bar{C} - C)^2 \right\rangle \approx \frac{2(2W+1)}{N} C^2$$

- here 4-point $\langle(a^i - \bar{a})(a^j - \bar{a})(a^k - \bar{a})(a^l - \bar{a})\rangle$ approximated by disconnected part $\langle(a^i - \bar{a})(a^j - \bar{a})\rangle\langle(a^k - \bar{a})(a^l - \bar{a})\rangle$ + two more
- $2W$ is similar to $B$ and has to be chosen to compromise.....but now exp small systematic error

Relative errors of estimates for $\sigma^2$, $\sigma_F^2$ (error of the error!)

| Quantity | Binning | $\Gamma$ |
|---|---|---|
| sys + stat = | $\tau/B + \sqrt{2B/N}$ | $\exp(-W/\tau) + \sqrt{2(2W+1)/N}$ |
| sum minimal for | $B = \tau(2N/\tau)^{1/3}$ | $W \simeq \ln(N/\tau)\tau/2$ |
| value | $3/2(2N/\tau)^{-1/3} \propto N^{-1/3}$ | $\sqrt{2\ln(N/\tau)\tau/N} \propto N^{-1/2}$ |
| ratio sys/stat = | $1/2$ | $1/\ln(N/\tau)$ |


0.6
0.5
0.4
0.3
0.2
0.1
0
total relative error of the error
Bins
Γ
10^1
10^2
10^3
10^4
10^5
N/τ

# Generalization: many variables

$$\bar{F} = f(\bar{a}_1, \bar{a}_2, \ldots), \quad \sigma_F^2 = \langle (\bar{F} - F)^2 \rangle = \frac{\bar{C}_F}{N}$$

$$\bar{C}_F = \bar{\Gamma}_F(0) + 2 \sum_{t=1}^{W} \bar{\Gamma}_F(t)$$

$$\bar{\Gamma}_F(t) = \frac{1}{N-t} \sum_{i=1}^{N-t} \sum_{\alpha\beta} \frac{\partial f}{\partial A_\alpha} \frac{\partial f}{\partial A_\beta} (a_\alpha^i - \bar{a}_\alpha)(a_\beta^{i+t} - \bar{a}_\beta)$$

a possible strategy is to

- compute: $\bar{a}_\alpha$
- evaluate $f_\alpha = \partial f / \partial A_\alpha$ *at argument* $\bar{a}_\alpha$
- "project" primary data: $a_F^i = \sum_\alpha f_\alpha a_\alpha^i$ to compute $\bar{C}, \bar{\Gamma}_F$ etc.
- then everything like primary variables...
- without signifact extra errors $f_\alpha$ may be taken as symmetric difference in $f$

# Generalization: Replica simulations

- $R$ independent simulations with the same algorithm, all equilibrated
- $\{a_\alpha^{i,r}\}, i = 1, \ldots, N_r, r = 1, \ldots, R$, total length $N = \sum_{r=1}^{R} N_r$
- may be made by dividing up one long run (neglecting edge effects)
- joint analysis will stabilize error estimates

$$\bar{a}_\alpha^r = \frac{1}{N_r} \sum_{i=1}^{N_r} a_\alpha^{i,r}, \quad \bar{\bar{a}}_\alpha = \frac{1}{N} \sum_{r=1}^{R} N_r \bar{a}_\alpha^r$$

$$\bar{\bar{\Gamma}}_{\alpha\beta}(t) = \frac{1}{N - Rt} \sum_{r=1}^{R} \sum_{i=1}^{N_r - t} (a_\alpha^{i,r} - \bar{\bar{a}}_\alpha)(a_\beta^{i+t,r} - \bar{\bar{a}}_\beta)$$

- avoid cross terms from different replica which contribute noise but no signal
- they are there, if only means and errors of replica are combined in the usual way

# Bonus material: unbias

two ways of estimating $F$:

$$\bar{\bar{F}} = f(\bar{\bar{a}}_\alpha), \quad \bar{F} = \frac{1}{N}\sum_{r=1}^{R} N_r\, f(\bar{a}^r_\alpha).$$

one can show $\langle \bar{\bar{F}} - F \rangle \simeq \frac{1}{R-1} \langle \bar{F} - \bar{\bar{F}} \rangle$ and eliminate the leading bias by replacing

$$\bar{\bar{F}} \to \frac{R\bar{\bar{F}} - \bar{F}}{R-1}$$

# Bonus material: $Q$-test

- expect $\{f(\bar{a}_\alpha^r)\}$ Gaussian with mean $F$ and width $\left\langle \left(f(\bar{a}_\alpha^r) - F\right)^2 \right\rangle = \frac{C_F}{Nr}$
- "fit to a constant" $K$ by minimizing

$$\chi^2(K) = \sum_r \frac{\left(f(\bar{a}_\alpha^r) - K\right)^2}{C_F/N_r}$$

- minimum at $K = \bar{F}$
- extra information:is $\chi^2$ reasonable?
- better: $Q = 1 - P(\chi^2/2, (R-1)/2)$ with $P =$ incomplete $\Gamma$-function
- $Q =$ probability to find $\chi^2$ higher or equal to the present one
- in practice: $Q \geq 0.1$ okay, $Q =$tiny $\Rightarrow$ our error estimate sucks (inconsistent)
- for example: all $N_r \ll \tau$ in replica with different initializations (hot, cold..)

# Matlab implementation

download MATLAB routine `UWerr.m` from `www-com.physik.hu-berlin.de/ALPHAsoft`

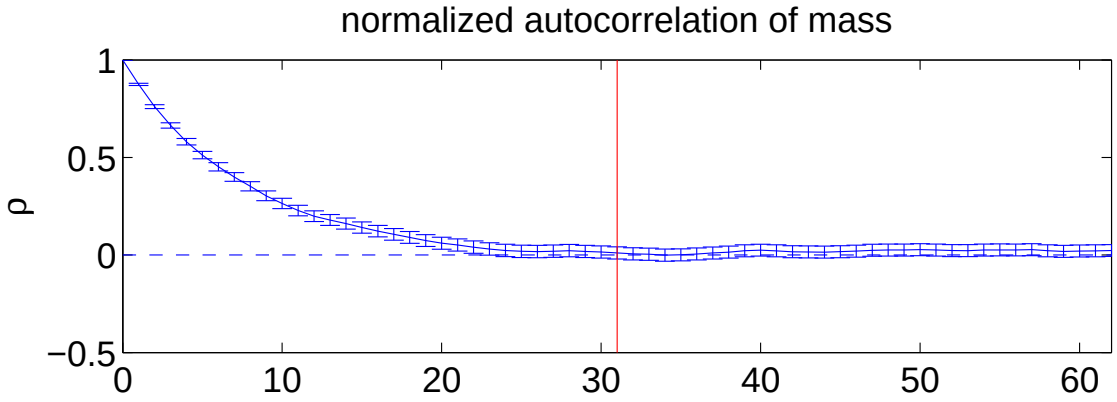
some features:

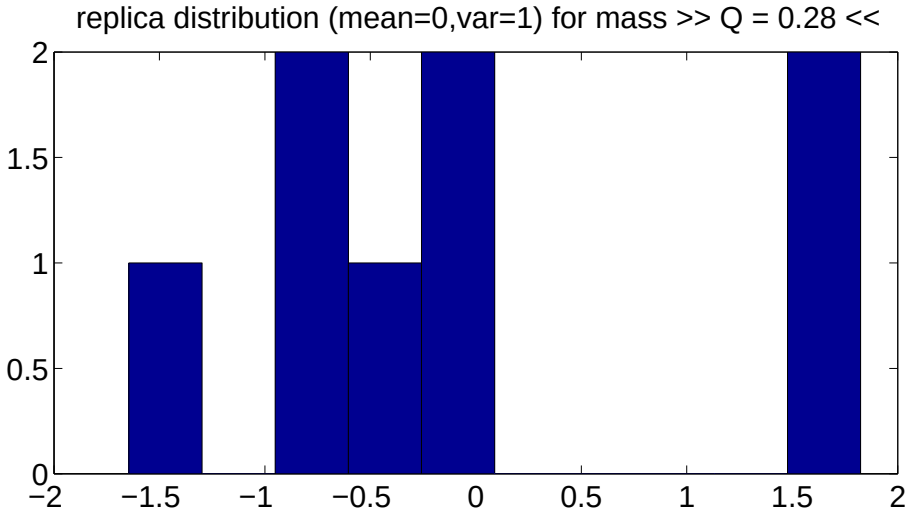
- estimate $f(A_1, A_2, A_3, \ldots)$ for user-defined routine for $f$ or for primary $A_\alpha$ themselves
- handle replica-simulations of arbitrary lengths, combined error analysis, check of statistical compatibility (goodnes of "fit" to a constant $Q$ computed)
- optimal window $W$ assessed semi-automatically; it needs a rough estimate of $S = \tau/\tau_{\mathrm{int},F}$
- plots generated for $\rho(t) = \Gamma(t)/\Gamma(0), \tau_{\mathrm{int}},$ replica distribution (see sample applications below)

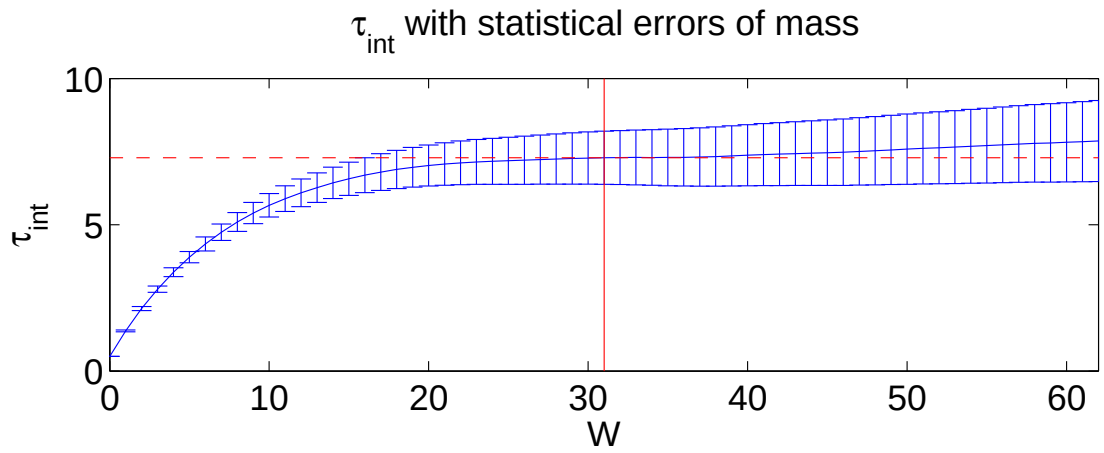
# Application 1: Synthetic data

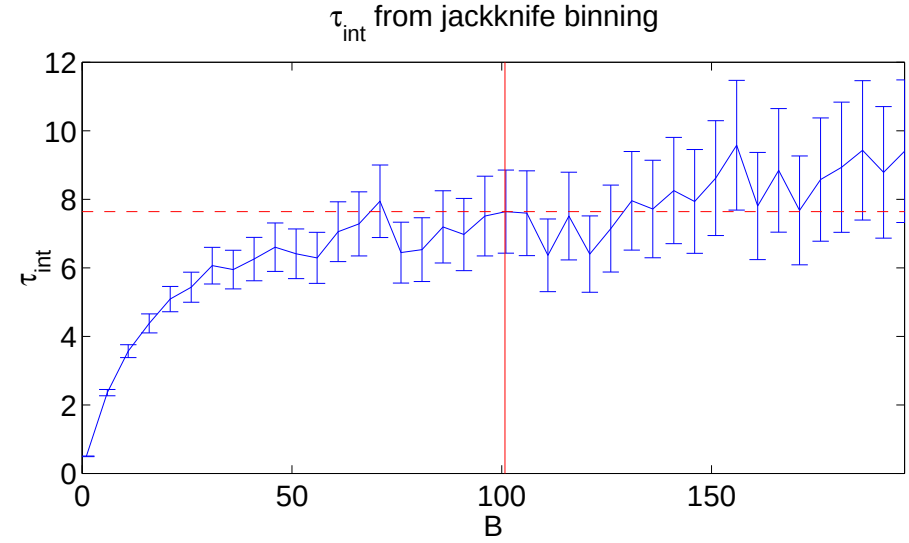
- artificial noisy autocorrelated data for 'time-slice correlation function' $A_1 = G(z)$, $A_2 = G(z+1)$
- measure exactly known $m = f(A_1, A_2) = \ln(A_1/A_2) = 0.2$
- noise (auto)correlated between $a_1^i$, $a_2^i$ (with $\tau = 4, 8$ components) such that $\tau_{\mathrm{int},m} = 7.92$
- 8 replica, 1000 measurements each, $\sigma_m = 0.0142$

```
>> Nr=[1000 1000 1000 1000 1000 1000 1000 1000]; S=1;
>> [value,dvalue,ddvalue,tauint,dtauint,Q]= ...
   UWerr(Data,S,Nr,'mass',@effmass,1,2);
>> [value,dvalue,ddvalue,tauint,dtauint,Q]

ans =  0.2128    0.0134    0.0008    7.2909    0.8021    0.2827

exact: 0.2       0.0142              7.92
```


normalized autocorrelation of mass
ρ
1
0.5
0
−0.5
0
10
20
30
40
50
60


replica distribution (mean=0,var=1) for mass >> Q = 0.28 <<
2
1.5
1
0.5
0
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2


τint with statistical errors of mass
τint
10
5
0
0
10
20
30
40
50
60
W

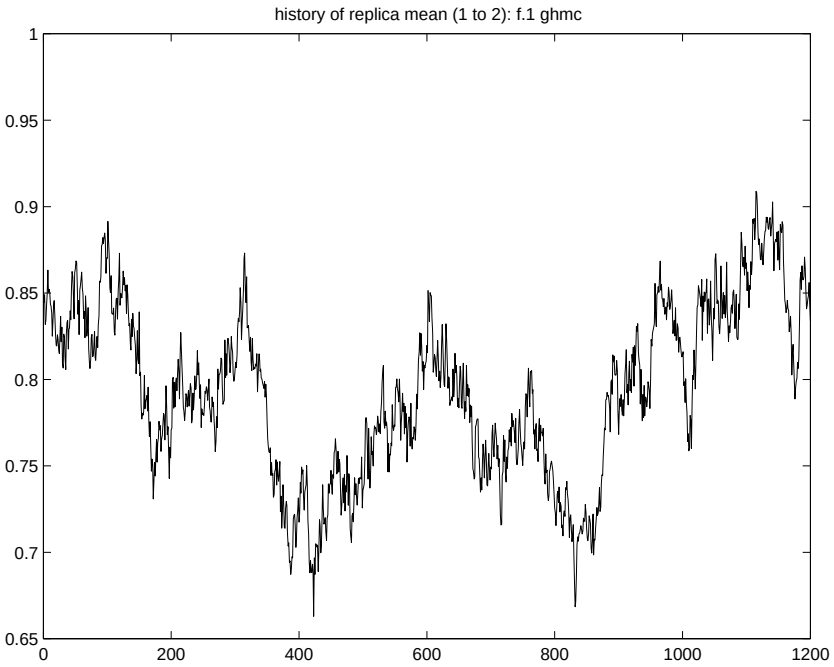

τint from jackknife binning
τint
12
10
8
6
4
2
0
0
50
100
150
B

# $Z_P$ with Dynamical fermions

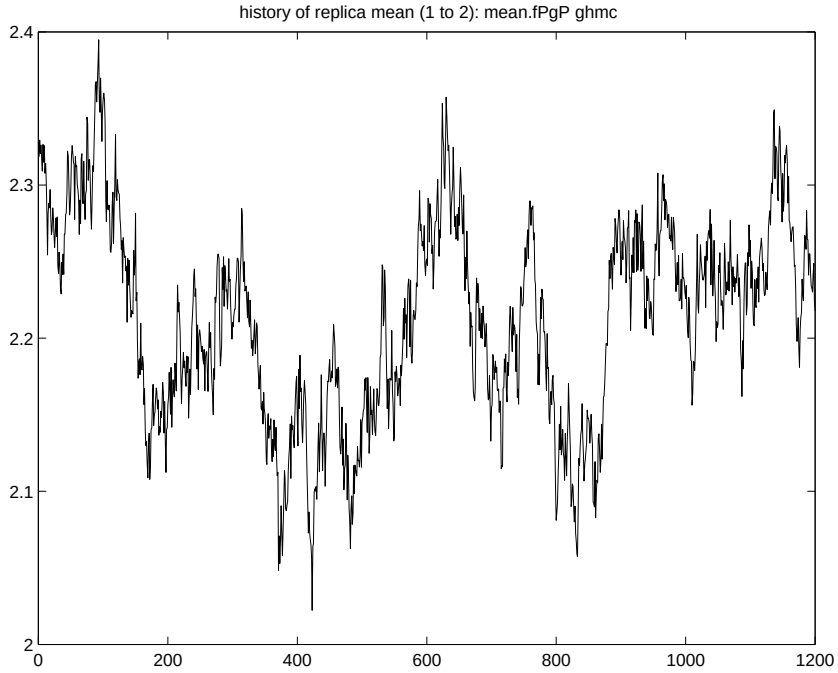
From the ALPHA running mass project with two lite flavors, a typical difficult run:
HMC with 2 pseudofermions, $L = 24, \beta = 8.02599, \kappa = 0.133063, \theta = 0.5$

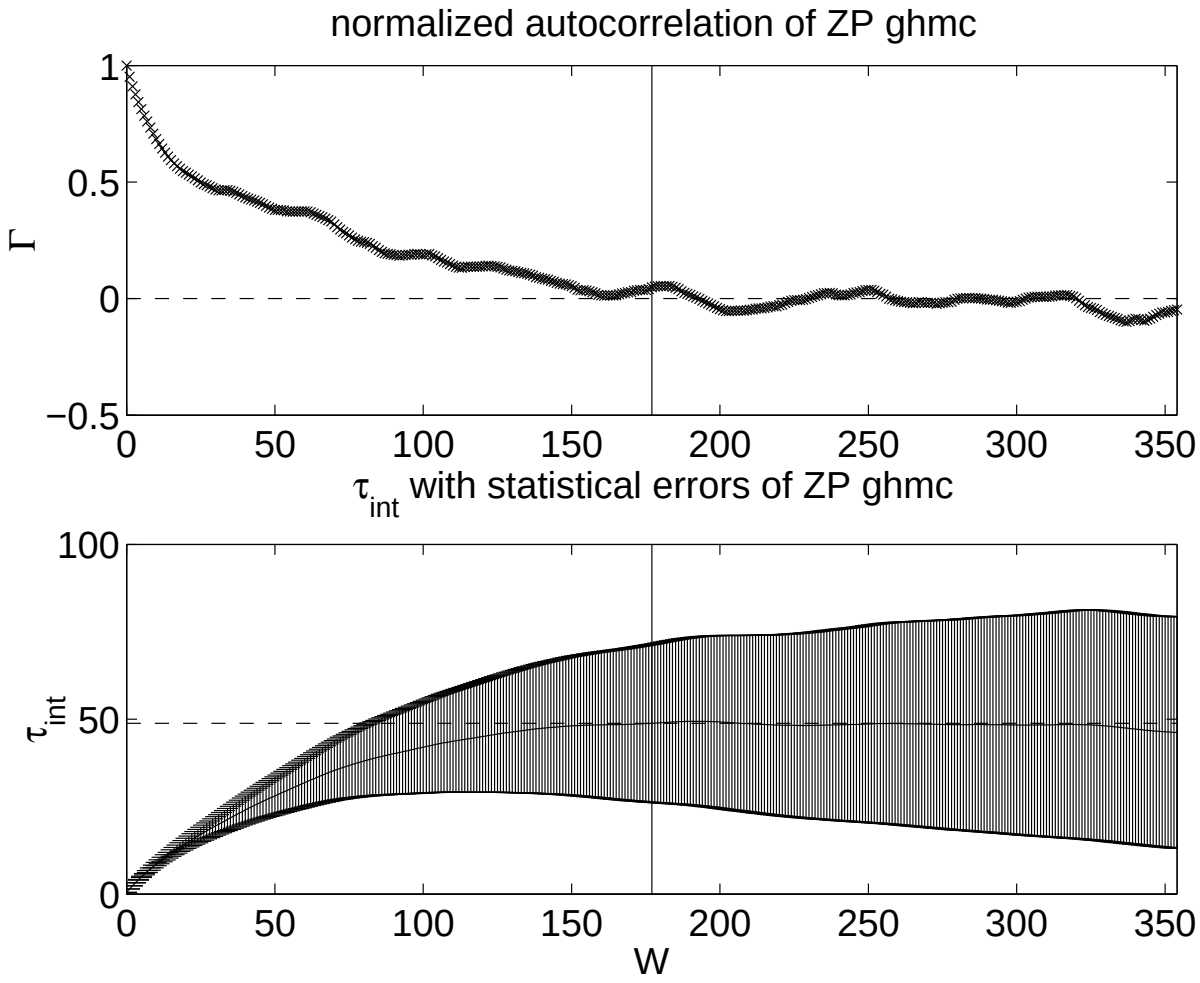
$$Z_P(L) = \frac{\sqrt{3 f_1}}{f_P(L)}, \quad f_1, f_P \text{ are SF correlations of } \bar{\psi}\gamma_5\psi, \quad L = \mu^{-1} = \text{ running-scale}$$

history $f_1$



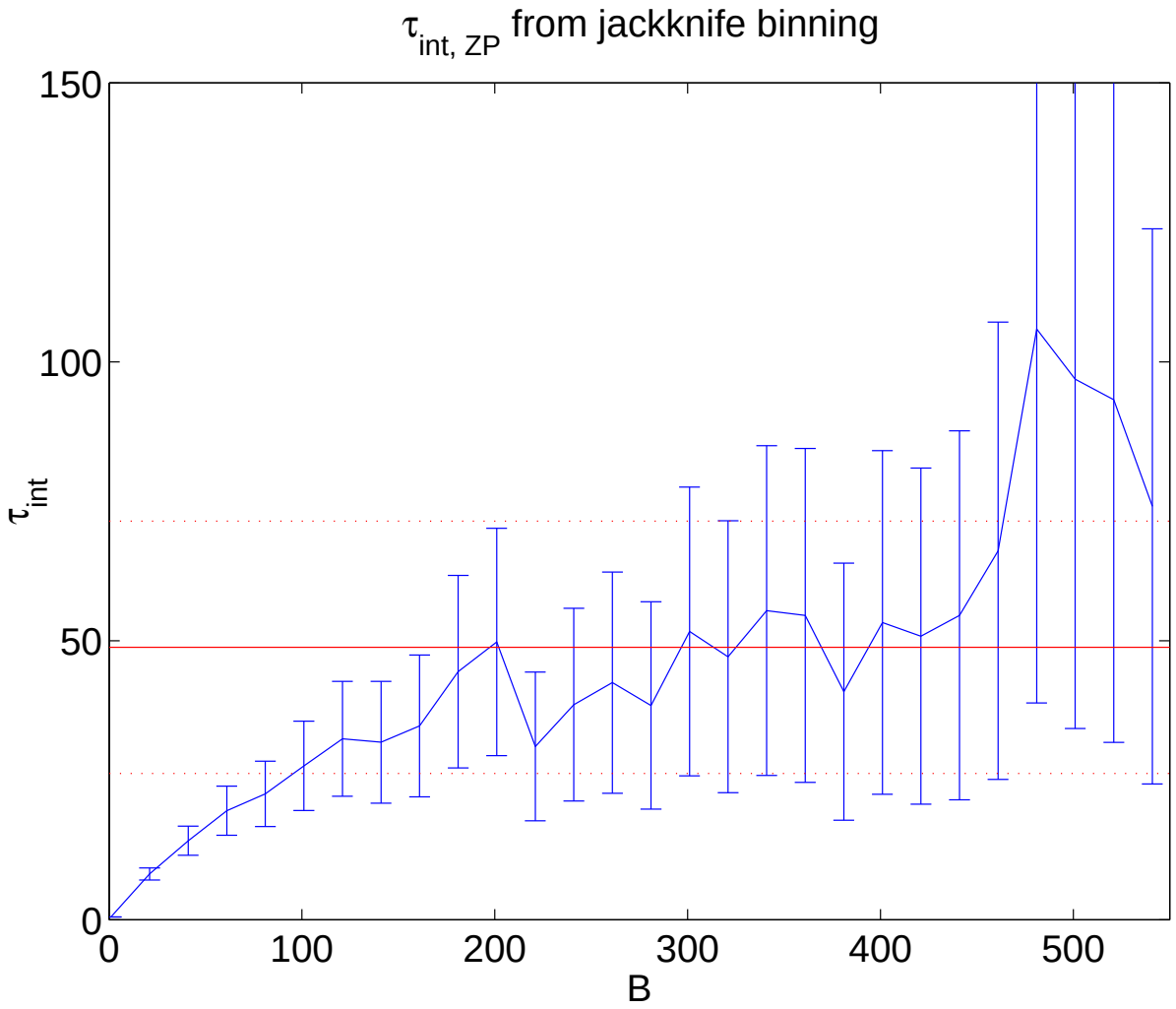

history $f_P$






```
 OBSERVABLE:   f1
  2 x 1200 measurements
  mean-value = 0.79325
       error = 0.01838
     tau_int =  88.92
error(Sokal) =  46.29
Q-value      =   0.80

 OBSERVABLE:   ZP
  2 x 1200 measurements
  mean-value = 0.69676
       error = 0.00322
     tau_int =  48.84
error(Sokal) =  22.62
Q-value      =   0.37
```

# Jackknife for $Z_P$



# Correlation $f_1$